УДК 004.89

МЕТОД АВТОМАТИЧЕСКОГО СЖАТИЯ СВЕРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ УСКОРЕНИЯ ИНФЕРЕНСА НА МОБИЛЬНЫХ УСТРОЙСТВАХ

Тарасевич Н.С. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Ефимова В.А. (ИТМО)

Введение. Сверточные нейронные сети (CNN) демонстрируют высокую точность в задачах компьютерного зрения (классификация, детекция, сегментация), однако их практическое применение на мобильных устройствах ограничено вычислительной сложностью и размером моделей. Это приводит к высокой задержке инференса, повышенному потреблению памяти и энергии, что критично для мобильных сценариев, включая распознавание объектов, AR/VR и навигацию. Существующие подходы к ускорению мобильного инференса включают проектирование эффективных архитектур и методы компрессии обученных моделей [1]: структурный прунинг каналов, квантование (включая пост-тренировочное квантование (PTQ) и квантование с обучением (QAT)) [2], а также дистилляцию знаний [3]. Несмотря на развитость перечисленных направлений, ключевой научно-практический разрыв состоит в том, что одинаковая степень «сжатия» по параметрам или FLOPs не гарантирует ускорение на конкретном устройстве – итоговая задержка зависит от рантайма, поддержки целочисленных вычислений, наличия оптимизированных ядер и структуры вычислительного графа. Поэтому актуальна задача автоматизированного выбора и комбинации методов сжатия, ориентированных на измеряемую задержку на целевом мобильном устройстве.

Основная часть. Предлагается модуль автоматического сжатия CNN под заданные ограничения мобильного устройства (latency/память) с формированием модели, готовой к запуску в мобильном рантайме. Подход объединяет несколько стратегий, сравниваемых по совокупности метрик (качество, размер, задержка), и выбирает оптимальную конфигурацию по компромиссу «качество–скорость–размер».

Ключевым режимом модуля является метод — сжатие CNN с учетом особенностей целевого оборудования, оптимизирующая не абстрактные показатели, а приближение реальной задержки на устройстве. В рамках метода осуществляются:

1. структурный прунинг по каналам для физического уменьшения модели;
2. селективная разреженность только в 1×1-свертках — там, где на практике чаще существуют эффективные оптимизированные ядра и разреженность потенциально приводит к ускорению;
3. INT8-ориентированное обучение (QAT) для стабильного целочисленного инференса;
4. дистилляция знаний от учителя (предобученной модели) для удержания точности при агрессивной компрессии;
5. автоматический выбор реализации для каждого 1×1-слоя (плотная, разреженная или низкоранговая) на основе аппаратно-зависимой оценки стоимости (LUT/калиброванные замеры) и заданного ограничения на задержку инференса.

Таким образом, предлагается не последовательное применение отдельных техник (например, «сначала прунинг, затем квантование»), а согласованная оптимизация структуры сети и формата вычислений с учетом особенностей мобильного исполнения. Это снижает риск ситуаций, когда формальная компрессия не дает ускорения, и позволяет получать практично быстрые модели при контролируемой потере качества.

Выводы. Предложен модуль автоматического сжатия CNN для мобильного инференса, ориентированный на измеряемые ограничения устройства и поддерживаемые режимы вычислений. Метод обеспечивает совместную компрессию (прунинг + селективная

разреженность + QAT + дистилляция знаний) и аппаратно-ориентированный выбор конфигурации 1×1-слоев, что позволяет улучшать баланс «качество–задержка–размер» в сравнении с изолированными или строго последовательными техниками. Практическое внедрение предполагает испытание на стандартных датасетах (например, CIFAR-10 и подмножество ImageNet) и обязательную валидацию задержки непосредственно на Android-устройстве в целевом рантайме. Результатом является библиотека/пайплайн, который по заданному ограничению на задержку инференса и ограничениям памяти автоматически подбирает конфигурацию сжатия и генерирует модель для мобильного приложения.

Список использованных источников:

1. Han S., Mao H., Dally W.J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding // ICLR. – 2016.
2. Jacob B. et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference // CVPR. – 2018.
3. Hinton G. et al. Distilling the Knowledge in a Neural Network // NeurIPS. – 2015.